Gated2Gated: Self-Supervised Depth Estimation from Gated Images Supplementary Document

Amanpreet Walia*[1,3] Stefanie Walz*[2] Mario Bijelic[4] Fahim Mannan[1]
Frank Julca-Aguilar[1] Michael Langer[3] Werner Ritter[2] Felix Heide[1,4]

[1]Algolux [2]Mercedes-Benz AG [3]McGill University [4]Princeton University

1. Introduction

This Supplemental Document provides additional information in support of the findings in the main manuscript. In Section 2, additional details on gated imaging are presented. Section 3 provides additional information about the temporal gated imaging dataset, and Section 4 describes further details of the network architecture. In Section 5, utilized masks and loss functions are explained in more detail and Section 6 provides additional evaluation details for the reference methods. Section 7 provides additional detail on the mask hyperparameters. In Section 8 and 9, we provide quantitative and qualitative results for our Gated2Gated framework.

2. Gated Imaging

In this section, we describe the gated imaging process in more detail than provided in the main manuscript. We assume a rectangular exposure function $p(t)$ and rectangular gating function $g(t)$. Considering a single pixel, which captures reflected photons of a point at a certain distance r, the corresponding photons require a round-trip time of $\frac{2r}{c}$ to reach the camera after being emitted by the source. This means we receive the signal $p\left(t-\frac{2r}{c}\right)$ at the sensor. The shutter of the sensor opens after a delay of ξ and remains open for the gate duration t_G. During the gating time t_G, all incident photons get integrated on the CMOS sensor. As such, the intensity value $Z(r)$ of the considered pixel is defined by the convolution of the gate pulse $g\left(t-\xi\right)$ and laser pulse $p\left(t-\frac{2r}{c}\right)$, that is

$$Z\left(r\right)=\phi\iota\int_{-\infty}^{\infty}g(t-\xi)p\left(t-\frac{2r}{c}\right)\beta(r)\mathrm{d}t, \tag{1}$$

where ϕ denotes the reflectivity, and the laser illumination ι defines the maximum amplitude of the laser pulse. The reflectivity ϕ depends on the spectral distribution of the scene illumination, the reflectance of the scene surfaces, and the atmosphere's water vapor content. Atmospheric effects, which are independent of object surfaces, are modeled by

$$\beta(r)=\frac{P_{\text{laser}}\tau_{\text{optics}}}{4\pi r^2\tan\left(\frac{\theta_H}{2}\right)\tan\left(\frac{\theta_V}{2}\right)}\frac{\rho^2}{F_{\text{num}}^2}\frac{\lambda}{hc}e^{-2\gamma r}, \tag{2}$$

with laser power P_{laser}, horizontal/vertical field of illumination θ_H/θ_V, pixel pitch ρ, aperture F_{num}, wavelength λ, Planck constant h, optical transmission τ_{optics}, and atmospheric attenuation coefficient γ. When ambient light occurs due to sunlight or other light sources in the scene, Equation 1 becomes

$$\begin{aligned}Z\left(r\right)&=\underbrace{\phi\iota}_{=\alpha}\int_{-\infty}^{\infty}g(t-\xi)p\left(t-\frac{2r}{c}\right)\beta(r)\mathrm{d}t+\underbrace{\phi\kappa\int_{-\infty}^{\infty}g(t-\xi)\mathrm{d}t}_{=\Lambda}\\&=\alpha C(r)+\Lambda,\end{aligned} \tag{3}$$

where κ denotes the ambient light falling on the considered point and $\phi\kappa$ indicates the level of reflected light reaching the sensor. Assuming constant ambient light during the gating time t_G, the captured ambient light results in $\Lambda=\phi\kappa\int_{-\infty}^{\infty}g(t-$

Laser		
Laser Power	P_{laser}	500 W
Wavelength	λ	808 nm
Horizontal Field of Illumination	θ_H	24°
Vertical Field of Illumination	θ_V	8°
Camera		
Pixel pitch	ρ	10 µm
Aperture	F_{num}	1.2
Optical transmission	τ_{optics}	0.64
Focal length	f	23 mm
Horizontal Field of View	θ_H	31.1°
Vertical Field of View	θ_V	17.8°
Resolution		1280x720

Table 1. Laser and camera specifications of the BrightwayVision BrightEye

$t_D)dt$. After read-out, the final measurement $Z\,(r)$ for each pixel location is obtained by

$$Z(r) \;=\; \alpha\, C(r) + \;\Lambda\; + \;\eta_g + \;\eta_p = \;\alpha\, C(r) + \;\hat{\Lambda}, \tag{4}$$

where η_p models the signal-dependent Poisson photon shot noise and η_g Gaussian read-out noise [3]. To increase the SNR, multiple laser pulses are integrated on the sensor before read-out.

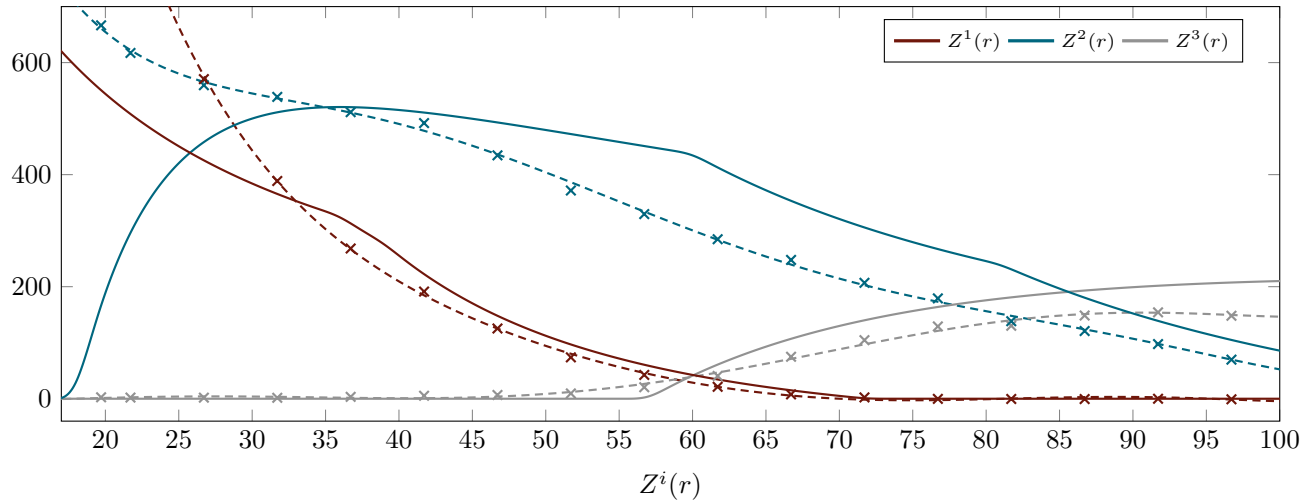


Figure 1. Measurements of the range-intensity-profiles used in this work. Solid lines identify the theoretically calculated profiles with the parameters from Table 1 und Table 2. Crosses mark real world measurements on reflectivity targets and their corresponding continuous Chebychev approximations are plotted with dashed lines.

In this work, we use Equation 4 to model gated images and to learn depth in self-supervised fashion by separating input gated images into albedo α, ambient illumination $\hat{\Lambda}$, and depth r. The gating parameters of the three input gated images are defined in Table 2 and the specifications for laser and camera employed in our work are reported in Table 1.

The range-intensity profiles resulting from the proposed gate parameters are documented in Figure 1. In addition to the analytically calculated profiles (full lines), Figure 1 lists real measured pixel intensity values (crosses) and the corresponding approximations (dashed lines), which are used for the cycle reconstruction approach. The real profiles are measured experimentally on targets with defined reflectivity at night and are approximated with Chebyshev polynomials T_n

$$T_0 = 1, \quad T_1 = x, \quad T_{n+1} = 2xT_n - T_{n-1}, \tag{5}$$

up to order of $N = 6$. We use the real-world measurements to calibrate unknown parameters such as the dark level of the camera or delays due to the signal runtime to adjust the real-world measurements to the analytical solutions. The gating

	Laser duration	Gate duration	Delay ξ	Pulses
Slice 1	240 ns	220 ns	260 ns	202
Slice 2	280 ns	420 ns	400 ns	591
Slice 3	370 ns	420 ns	750 ns	770

Table 2. Definitions of the gating parameters that we use for the experimental acquisition in this work.

settings correspond to depth slices between 3-72 m, 18-123 m, and 57-176 m, respectively. Since the laser illumination decreases quadratic with the distance, we compensate for the lower illumination with a higher number of laser pulses for far distances.

3. Temporal Gated Imaging Dataset

In this section, we provide additional details on the gated video sequences that were captured to train the proposed method. Two temporal sequence examples are shown in Figure 3. Note that existing gated datasets [1] do not include temporal sequences. The dataset consists of 1,835 video sequences captured at 10 Hz. The videos were uniformly sampled at 0.1 Hz to extract an initial set of keyframes. We then selected the top 13,000 most interesting keyframes and extracted a long sequence (then at 10 Hz) centered around those keyframes. The final key frames include a wide variety of scenarios, including day, night, fog, snow, and clear weather conditions. The distributions for different times of day and weather conditions are reported in Figure 2. The complete dataset, which includes synchronized monocular gated and RGB images, adds up to 200 TB.

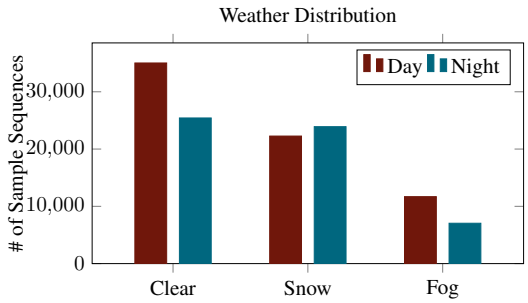


Figure 2. Distribution of different adverse weather effects within the captured temporal gated dataset.

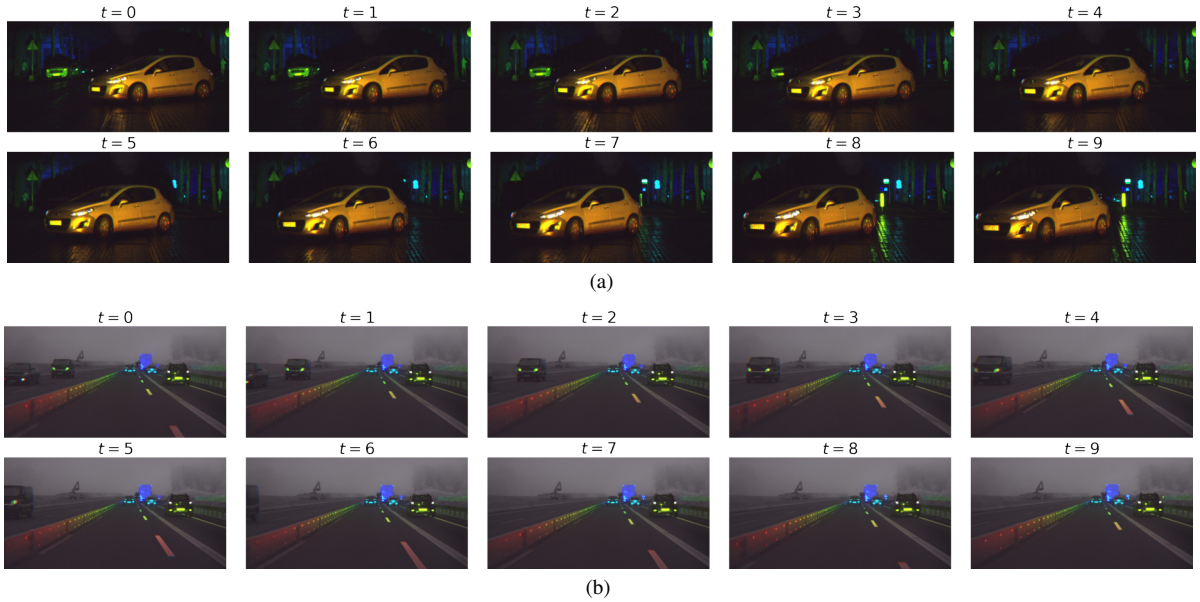


Figure 3. Sample frames from a video sequence in (a) night and (b) light fog weather conditions. Each frame contains three gated slices, concatenated as R,G,B channels for visualization (red color correspond to close depth slices, and blue correspond to far depth slices).

4. Additional Network Details

The Gated2Gated architecture consists of three different components: a depth prediction network $f_{\mathbf{r}}$, a pose prediction network $f_{\mathbf{t}\to\mathbf{n}}$, and an albedo/ambient prediction network $f_{\mathbf{\Lambda}\alpha}$. For the depth prediction network, we adopt the PackNet architecture [10] with four 3D-convolutions, and for pose prediction, we use the network proposed in [13]. The $f_{\mathbf{\Lambda}\alpha}$ network consists of one encoder and two separate decoders for albedo and ambient illumination. The encoder and decoder networks are a variant of the popular U-Net. The encoder consists of four convolutional layers with 3×3 kernels and a max-pooling operation and batch normalization after each layer. The decoder includes four pairs of flat and transposed convolutional layers. Furthermore, we use skip connections between the encoder and decoder to share semantic context at different feature scales. The detailed architecture descriptions of our modified U-Net-based network for joint albedo and ambient estimation is provided in Table 3.

ENCODER			
Layers #	**Layer Description**		**Output Shape**
0	Input		3 x H x W
1a	ConvBlock-1	Conv 3×3	$32 \times H \times W$
		LeakyReLU ($\alpha = 0.2$)	$32 \times H \times W$
		BatchNorm2D	$32 \times H \times W$
		Conv 3×3	$32 \times H \times W$
		LeakyReLU ($\alpha = 0.2$)	$32 \times H \times W$
		BatchNorm2D	$32 \times H \times W$
1b	MaxPool2D		$32 \times \frac{H}{2} \times \frac{W}{2}$
2a	ConvBlock-2		$64 \times \frac{H}{2} \times \frac{W}{2}$
2b	MaxPool2D		$64 \times \frac{H}{4} \times \frac{W}{4}$
3a	ConvBlock-3		$128 \times \frac{H}{4} \times \frac{W}{4}$
3b	MaxPool2D		$128 \times \frac{H}{8} \times \frac{W}{8}$
4a	ConvBlock-4		$256 \times \frac{H}{8} \times \frac{W}{8}$
4b	MaxPool2D		$256 \times \frac{H}{16} \times \frac{W}{16}$
5	ConvBlock-5		$512 \times \frac{H}{16} \times \frac{W}{16}$

ALBEDO (DECODER)				AMBIENT (DECODER)			
Layer #	**Layer Description**		**Output Shape**	**Layer Description**			**Output Shape**
6a	Upsampling-1	ConvTranspose2D (kernel = 2)	$256 \times \frac{H}{8} \times \frac{W}{8}$	Upsampling-1	ConvTranspose2D (kernel = 2)		$256 \times \frac{H}{8} \times \frac{W}{8}$
		BatchNorm2D			BatchNorm2D		
6b	Concat-1	Layer #6a $\oplus$ Layer #4a	$512 \times \frac{H}{8} \times \frac{W}{8}$	Concat-1	Layer #6a $\oplus$ Layer #4a		$512 \times \frac{H}{8} \times \frac{W}{8}$
6c	UpConvBlock-1	Conv (3x3)	$256 \times \frac{H}{8} \times \frac{W}{8}$	UpConvBlock-1	Conv (3x3)		$256 \times \frac{H}{8} \times \frac{W}{8}$
		LeakyReLU ($\alpha = 0.2$)			LeakyReLU ($\alpha = 0.2$)		
		BatchNorm2D			BatchNorm2D		
		Conv (3x3)			Conv (3x3)		
		LeakyReLU ($\alpha = 0.2$)			LeakyReLU ($\alpha = 0.2$)		
		BatchNorm2D			BatchNorm2D		
7a	Upsampling-2		$128 \times \frac{H}{4} \times \frac{W}{4}$	Upsampling-2			$128 \times \frac{H}{4} \times \frac{W}{4}$
7b	Concat-2	Layer #7a $\oplus$ Layer #3a	$256 \times \frac{H}{4} \times \frac{W}{4}$	Concat-2	Layer #7a $\oplus$ Layer #3a		$256 \times \frac{H}{4} \times \frac{W}{4}$
7c	UpConvBlock-2		$128 \times \frac{H}{4} \times \frac{W}{4}$	UpConvBlock-2			$128 \times \frac{H}{4} \times \frac{W}{4}$
8a	Upsampling-3		$64 \times \frac{H}{2} \times \frac{W}{2}$	Upsampling-3			$64 \times \frac{H}{2} \times \frac{W}{2}$
8b	Concat-3	Layer #8a $\oplus$ Layer #2a	$128 \times \frac{H}{2} \times \frac{W}{2}$	Concat-3	Layer #8a $\oplus$ Layer #2a		$128 \times \frac{H}{2} \times \frac{W}{2}$
8c	UpConvBlock-3		$64 \times \frac{H}{2} \times \frac{W}{2}$	UpConvBlock-3			$64 \times \frac{H}{2} \times \frac{W}{2}$
9a	Upsampling-4		$32 \times H \times W$	Upsampling-4			$32 \times H \times W$
9b	Concat-4	Layer #9a $\oplus$ Layer #1a	$64 \times H \times W$	Concat-4	Layer #9a $\oplus$ Layer #1a		64 x Hx W
9c	UpConvBlock-3		$32 \times H \times W$	UpConvBlock-3			$32 \times H \times W$
10	Conv1D		$1 \times H \times W$	Conv1D			$1 \times H \times W$

Table 3. Our U-Net based architecture for $f_{\mathbf{\Lambda}\alpha}$. Here, $\oplus$ defines channel concatenation across feature tensors, α defines slope of LeakyReLU. We have two output heads : 1) Albedo and 2) Ambient sharing same encoder backbone. We have modified existing U-Net by adding BatchNorm2D layers, which leads to more stable training in our experiments.

5. Additional Loss Details

Next, provide additional information on the used loss functions and validity masks.

5.1. Additional Detail on Passive Supervision Loss

The ambient illumination can include sunlight and other light sources. By using a narrow-band filter adjusted to the wavelength of the laser, the gated camera is able to reduce the ambient light significantly. However, sunlight contains a strong NIR component, and ambient illumination is still present in the gated slices, especially during daytime. To learn the ambient illumination of the scene, our network gets supervised by a passive image captured by the same gated camera and deactivated laser illumination. The passive image $\tilde{\mathbf{Z}}_t^p$ is recorded with an optimized exposure duration to allow the best capturing of all image details even in dark conditions and strong blending sunlight. To align the passive exposure time with the active gated slices, the passive intensities have to be scaled accordingly with factor s_p, assuming a linear exposure curve. Additionally, we have to ensure equal passive illumination in each of the gated slices. Since the three gated slices are captured with a different number of laser pulses and varying gating duration (see Figure 2), we add to each slice an variable exposure with deactivated illumination to ensure the same total integration for each gated slice. This ensures that the passive part within each gated slice is equal. This facilitates the learning and supervision of the ambient component. The final passive supervision loss is defined by the photometric loss between ambient prediction $\hat{\mathbf{\Lambda}}$ and scaled passive image $\mathbf{Z}_t^p$, that is

$$\mathcal{L}_p(\hat{\mathbf{\Lambda}}, \mathbf{Z}_t^p) = 0.85 \cdot \frac{1-SSIM(\hat{\mathbf{\Lambda}}, s_p\tilde{\mathbf{Z}}_t^p)}{2} + 0.15 \cdot ||\hat{\mathbf{\Lambda}} - s_p\tilde{\mathbf{Z}}_t^p||_1. \tag{6}$$

5.2. Additional Qualitative Results for Cycle Reconstruction Predictions

New now take a closer look at the components required for the cycle reconstruction loss. Our method predicts scene depth, surface albedo (NIR reflectivity) and ambient illumination. The depth-dependent intensity profile, albedo and ambient components are combined in our method to reconstruct the input gated images.

Figure 4 shows qualitative examples of the three input gated images, the passive image, as well as the predicted depth, ambient illumination, and albedo. At night, the ambient images are almost completely dark. Only active light sources, such as car lights or traffic lights, are captured in the ambient image. During the daytime, sunlight results in strong ambient components, resulting in bright images. In contrast, the albedo images are consistent for day and night. Please note that shadows behind objects appear due to the different mounting positions of the active illumination source and the gated camera. Furthermore, it is possible to recognize the oval shape of the active illumination represented by the dark edges in the albedo.

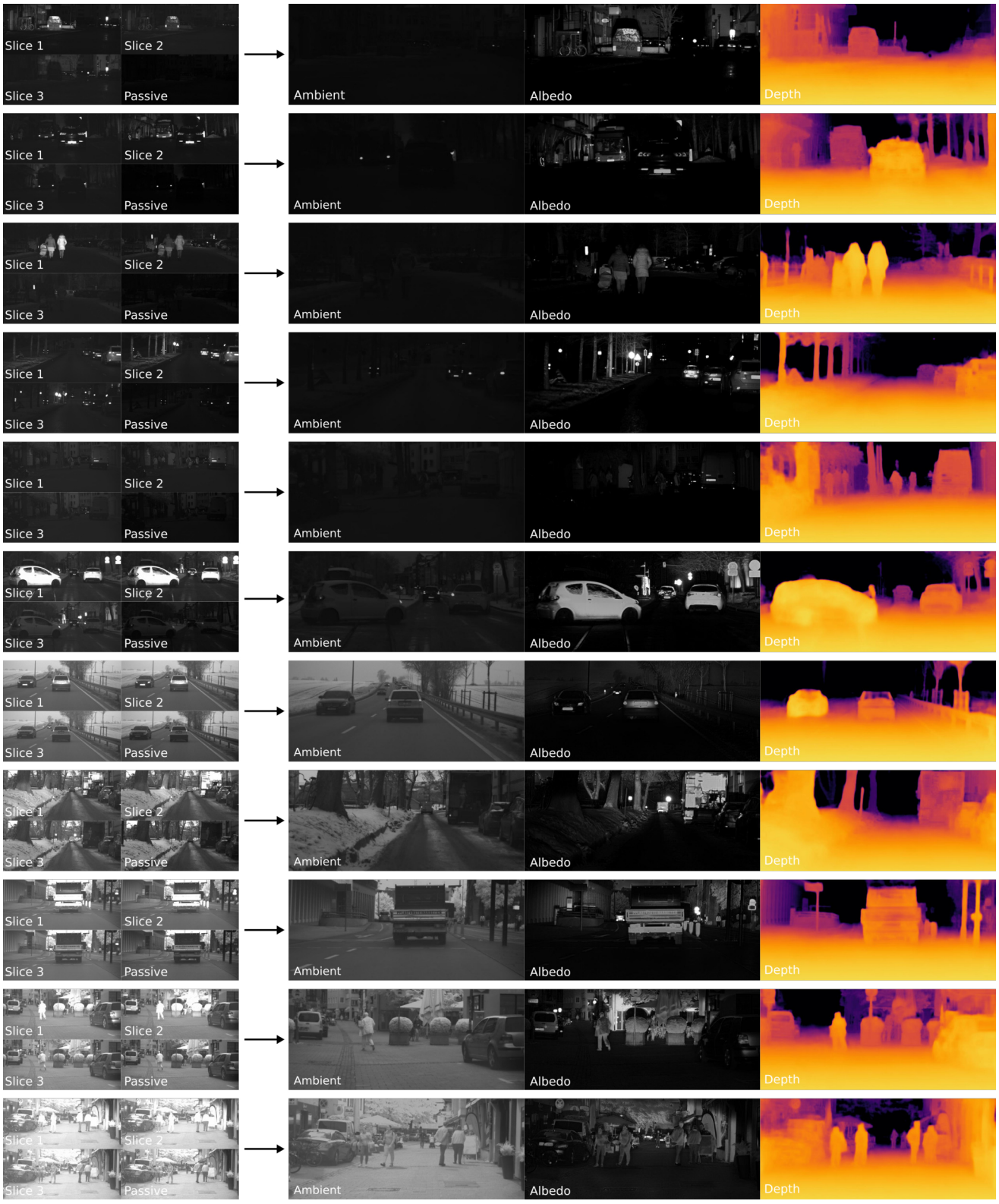


Figure 4. Qualitative examples of the three input gated slices and the passive images, and the predicted components. As a by-product, our method reconstructs albedo and ambient illumination of the scene in addition to the depth information. The ambient image captures the sunlight and activates light sources, such as vehicle lights or traffic lights, while the albedo captures the NIR reflectively of objects and the illumination from the active source.

5.3. Additional Explanations on Infinity Correction Mask

We next discuss the infinity correction masks. The masks are necessary to handle dynamic scene objects not captured by the rigid scene transformation between to adjacent temporal frames. Due to zero relative motion, dynamic objects might remain stationary with respect to the ego-vehicle.This corresponds to these objects projected at infinity distance. Hence, the depth is wrongly predicted. To prevent self-supervision with such incorrect depth cues, we employ infinity correction masks.

We rely on gated intensity cues to get rough depth estimates and identify regions out of bounds. The intensity depth cues can be seen in Figure 1. The ranges for the first slice is 3-72 m, for the second slice 18-123 m and for the last slice 57-176 m. The intensity relation allows to reason about the pixel distances depending on its intensity triplet from three different slices. Additionally, for each overlapping slice, there are intersecting distances, where one of the slices has a higher intensity. For example, from a distance of approx. 80 m, the last slice has the highest intensity. By demanding that intensities in the first and second slice are higher compared to the last slice we can filter all points closer than 80 m. By comparing the intensities between different gated slices we estimate areas up to the intersection points providing us a set of pixels with upper bound depth.

6. Additional Evaluation Details

Next, we provide additional details for the evaluation and training of the state-of-the-art methods we compare to for this work. To evaluate methods that use gated images as input, we crop 150 pixels on each side of the gated images, resulting in a final resolution of 420×980 pixels. This center crop is required due to the reduced laser illumination at the edges of the images where no modulation is present. For the evaluation of the RGB based methods, we crop the RGB images to a similar view and rescale them to ensure a fair comparison between the different modalities. For the training of the stereo-based self-supervised approach [4], we used the weights of the best model available as initialization and finetune them on the Gated2Depth training dataset [9]. Finetuning of Sparse-To-Dense [11] is not possible since neither dense nor semi-dense ground truth depth is available on our dataset. All self-supervised monocular approaches [6, 10] are initialized with the best RGB model available and are finetuned on the proposed temporal dataset.

7. Additional Details on Mask Hyperparameters

The cycle mask depends on the parameters γ and θ. While γ is used to remove saturated pixels, θ is used as a lower bound for the SNR value of the gated slices, restricting training to areas with reliable illumination cues. This is important as areas that are typically not illuminated have low SNR and saturated regions result in a non-linearity not handled by our albedo and ambient models. The infinity correction mask allows us to remove regions without temporal cues, e.g., regions that have zero relative velocity with the ego-vehicle. To this end, we compare the intensity values of the close range gated slices with the intensity values of the long range gated slice, and define a lower bound for the first (based on the second) as defined in Eqs. 16 & 17 of the main paper. This can also be seen as selecting pixels that have values exceeding a minimum floor in both the close and middle gated slices, which does not happen for dynamic objects unless they are moving at the speed of the ego-vehicle. As our goal is to select all pixels with relevant depth cues at either close or middle distances, without distinguishing between them (Eq. 18 in main paper), we only require one value of c for both the first and middle slices. Masks for different hyperparameter settings are shown in Fig. 5. We find suitable parameters with random search on the validation set. The results from this search are shown in Table 4. In the future, we envision learning optimal masks jointly with the model parameters.

	DAY					NIGHT				
c	0.7	0.995	0.995	0.995	1.5	0.7	0.995	0.995	0.995	1.5
θ	0.04	0.01	0.04	0.1	0.04	0.04	0.01	0.04	0.1	0.04
RMSE	14.24	14.28	**13.42**	15.1	14.27	11.94	11.94	**11.13**	13.0	12.49
MAE	11.01	10.48	**9.96**	11.94	10.90	8.63	8.17	**7.65**	9.53	8.90

Table 4. Effect of mask hyper-parameters on depth metrics [m].

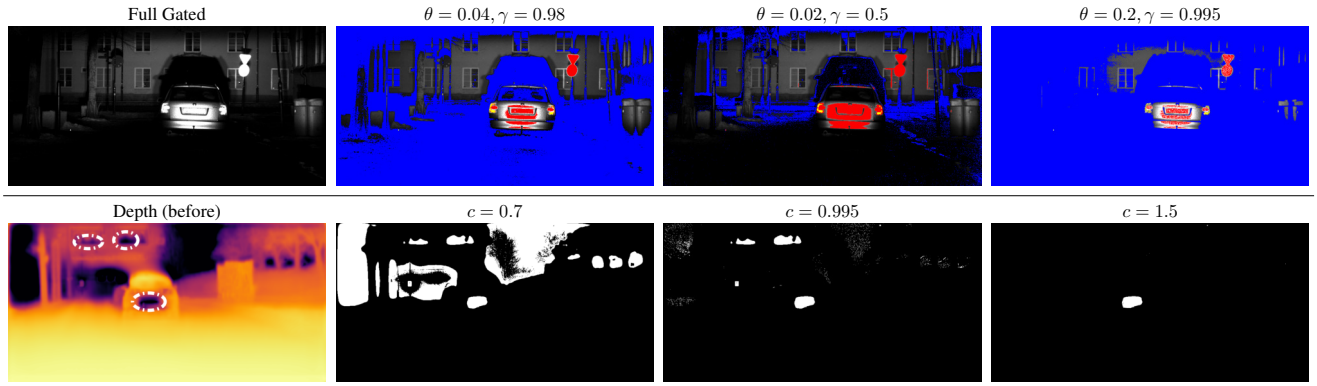


Figure 5. Cycle loss mask (top row) and infinity correction mask (bottom row) for infinite depth holes (circled) with different hyperparameters. For cyclic masks, red areas correspond to saturated pixels and blue areas to low SNR.

8. Additional Quantitative Result

In this section, we provide additional quantitative results. Specifically, in Subsection 8.1 we present ablation experiments, and in Subsection 8.2 we provide depth-resolved evaluations in different weather conditions.

8.1. Additional Ablation Studies

We investigate the influence of the depth network architecture, pretraining schemes, resolution, loss and mask combinations. All ablation experiments are performed on models trained till convergence, and the best performing epoch is used for evaluation. Table 5 shows the impact of using the cycle reconstruction mask and the infinity correction mask on the model performance. Since multipath effects and infinity holes constitute only a minor percentage of the Gated2Depth test dataset, we manually created a subset of this dataset containing only multipath and infinity holes. Evaluating on this new dataset helps illustrating the impact of the masks. The results in Table 5 validate, that the model trained without any mask performs worst. Adding one mask, already leads to a notable performance boost. However, using all masks together significantly increases the performance and results in a 14% improved MAE Metric. Table 6 lists further ablation experiments, evaluated on the full Gated2Depth test dataset. In this Table, we investigate different depth network architectures. Aside from the Resnet18 used in Monodepth2 [6], we also trained models with a Packnet [10] architecture consisting of four or 8 3D-convolutions. The results show that best performance is obtained with the Packnet architecture consisting of four 3D-convolutions. Furthermore, we found that passive supervision helps to stabilize the training process and models without this loss component often diverge. Table 6 demonstrates that a lower resolution leads to worse results than models trained with high resolution. Models trained without temporal loss and cycle loss only suffer from low SNR and saturated regions, which explains the moderate performance. Combining all loss functions and all proposed masks in the final model leads to a significant improvement in all metrics, validating the proposed model and loss choices.

	Method	Resolution	Depth Net	Pretrained	Temporal Loss	Cycle Loss	Passive Loss	Cycle Mask	Infinity Mask	RMSE	ARD	MAE	δ_1	δ_2	δ_3
	Evaluation on a Test Set with predominant Multipath and Infinity Hole Effects														
DAY	GATED2GATED w/o masks	512x1024	Packnet(D=4)	✓	✓	✓	✓	✗	✗	9.63	0.19	4.93	82.51	93.06	96.08
DAY	GATED2GATED with infinity mask	512x1024	Packnet(D=4)	✓	✓	✓	✓	✗	✓	9.14	0.17	4.82	83.10	93.86	96.60
DAY	GATED2GATED with cycle mask	512x1024	Packnet(D=4)	✓	✓	✓	✓	✓	✗	8.80	0.17	4.50	83.91	94.18	96.82
DAY	**GATED2GATED** (final)	512x1024	Packnet(D=4)	✓	✓	✓	✓	✓	✓	**8.59**	**0.15**	**4.24**	**85.86**	**94.48**	**96.85**
NIGHT	GATED2GATED w/o masks	512x1024	Packnet(D=4)	✓	✓	✓	✓	✗	✗	10.92	0.22	5.40	83.22	91.61	94.44
NIGHT	GATED2GATED with infinity mask	512x1024	Packnet(D=4)	✓	✓	✓	✓	✗	✓	9.98	0.18	4.97	85.13	93.58	95.85
NIGHT	GATED2GATED with cycle mask	512x1024	Packnet(D=4)	✓	✓	✓	✓	✓	✗	**9.87**	0.18	4.64	85.67	93.54	**95.86**
NIGHT	**GATED2GATED** (final)	512x1024	Packnet(D=4)	✓	✓	✓	✓	✓	✓	9.95	**0.17**	**4.62**	**86.80**	**93.61**	95.68

Table 5. Ablation studies evaluated on a subset of the Gated2Depth test set containing predominant multipath effects and infinity holes. By applying the proposed infinity correction mask and the cycle reconstruction mask the MAE metric is increased up to 14%.

8.2. Additional Depth-resolved Evaluations

Next, we provide additional information on the depth-resolved evaluation used to assess depth prediction performance in adverse weather scenarios in Table 2 of the main manuscript. Depth-resolved evaluations were introduced in [7]. In addition, we also analyze the ground truth LiDAR depth histogram of the STF dataset [1] in Figure 6. Here, the frequency of LiDAR ground truth points per distance is shown. It can be seen that there is an increase of points in close distance and a stronger

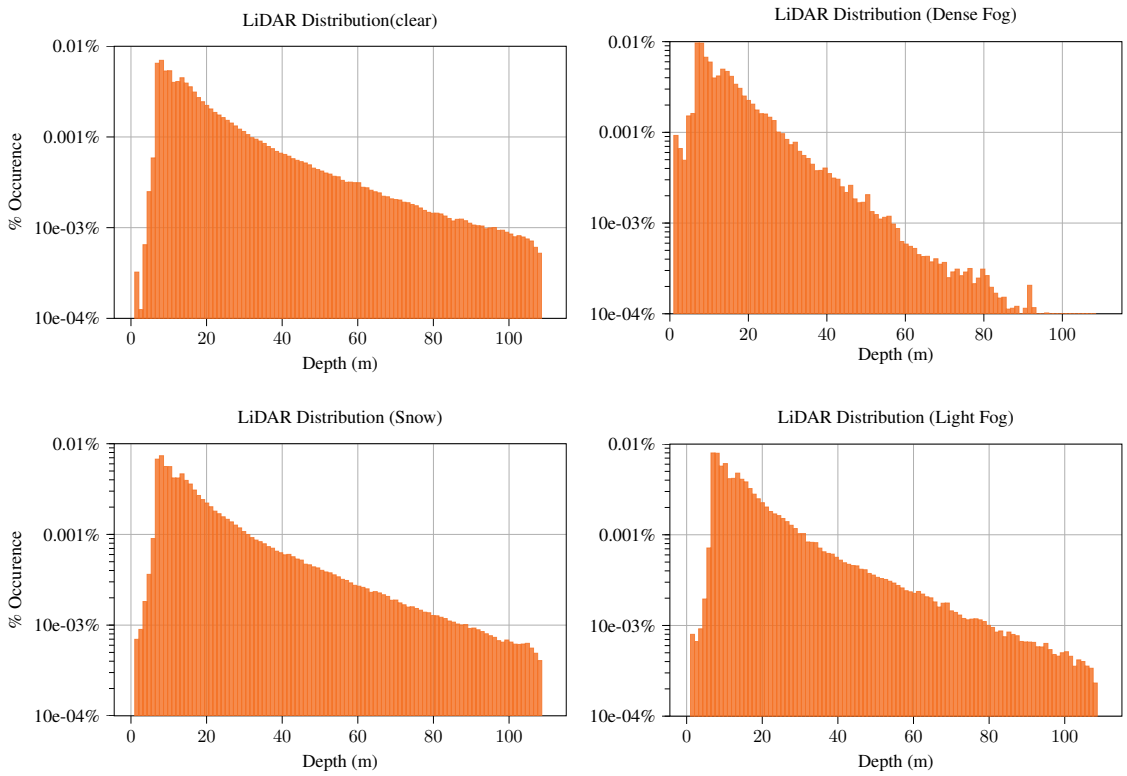


Figure 6. Distribution of LiDAR points in different adverse weather settings.

decay for further ranges in all weather settings compared to clear conditions. To handle this imbalance, we equally weigh the evaluation results for 7 m bins in the 3-80 m range. Model performance for distance bins are shown in Figure 7. As the Figure shows, at longer ranges and in day time conditions, our proposed Gated2Gated approach achieves the best performance. For

	Method	Resolution	Depth Net	Pretrained	Temporal Loss	Cycle Loss	Passive Loss	Cycle Mask	Infinity Mask	RMSE	ARD	MAE	δ_1	δ_2	δ_3
	Evaluation on the Gated2Depth Test Dataset														
DAY	BASELINE	512x1024	Packnet(D=4)	✗	✓	✗	✗	✗	✗	12.44	0.27	7.23	66.32	85.85	92.40
DAY	GATED2GATED low res.	256x512	Packnet(D=4)	✓	✓	✓	✗	✗	✗	20.34	0.67	15.17	25.42	50.34	70.71
DAY	GATED2GATED cycle only	512x1024	UNet	✗	✗	✓	✗	✗	✗	14.57	0.38	9.38	42.26	69.77	84.09
DAY	GATED2GATED Resnet18	512x1024	Resnet18	✓	✓	✓	✓	✓	✓	12.12	0.34	8.10	49.43	81.71	91.22
DAY	GATED2GATED Full Packnet	512x1024	Packnet(D=8)	✓	✓	✓	✓	✓	✓	9.73	0.22	5.13	81.30	91.97	95.32
DAY	GATED2GATED from scratch	512x1024	Packnet(D=4)	✗	✓	✓	✓	✓	✗	9.23	0.22	4.98	80.52	92.32	95.62
DAY	GATED2GATED from scratch	512x1024	Packnet(D=4)	✗	✓	✓	✓	✓	✓	10.15	0.25	5.60	77.26	90.89	94.81
DAY	GATED2GATED from scratch	512x1024	Packnet(D=4)	✗	✓	✓	✗	✓	✓	50.44	2.70	47.44	8.76	14.98	21.80
DAY	GATED2GATED w/o passive	512x1024	Packnet(D=4)	✓	✓	✓	✗	✓	✓	8.87	0.19	4.54	83.22	92.98	95.90
DAY	GATED2GATED w/o masks	512x1024	Packnet(D=4)	✓	✓	✓	✓	✗	✗	9.29	0.22	4.99	80.74	91.88	95.40
DAY	GATED2GATED with infinity mask	512x1024	Packnet(D=4)	✓	✓	✓	✓	✗	✓	9.14	0.20	4.99	80.82	92.26	95.58
DAY	GATED2GATED with cycle mask	512x1024	Packnet(D=4)	✓	✓	✓	✓	✓	✗	8.85	0.20	4.75	81.20	92.57	95.85
DAY	**GATED2GATED** (final)	512x1024	Packnet(D=4)	✓	✓	✓	✓	✓	✓	**8.46**	**0.17**	**4.37**	**83.56**	**93.12**	**96.09**
NIGHT	BASELINE	512x1024	Packnet(D=4)	✗	✓	✗	✗	✗	✗	12.15	0.27	6.87	69.14	86.93	92.57
NIGHT	GATED2GATED low res.	256x512	Packnet(D=4)	✓	✓	✓	✗	✗	✗	16.98	0.56	12.39	31.60	56.80	75.82
NIGHT	GATED2GATED cycle only	512x1024	UNet	✗	✗	✓	✗	✗	✗	17.25	0.61	11.78	36.06	57.20	69.82
NIGHT	GATED2GATED Resnet18	512x1024	Resnet18	✓	✓	✓	✓	✓	✓	11.80	0.29	7.30	57.16	83.24	89.92
NIGHT	GATED2GATED Full Packnet	512x1024	Packnet(D=8)	✓	✓	✓	✓	✓	✓	10.31	0.27	5.47	80.12	90.56	93.86
NIGHT	GATED2GATED from scratch	512x1024	Packnet(D=4)	✗	✓	✓	✓	✓	✗	10.37	0.28	5.55	78.86	90.74	93.99
NIGHT	GATED2GATED from scratch	512x1024	Packnet(D=4)	✗	✓	✓	✓	✓	✓	11.49	0.33	6.23	76.12	89.10	93.00
NIGHT	GATED2GATED from scratch	512x1024	Packnet(D=4)	✗	✓	✓	✗	✓	✓	51.33	2.89	48.27	8.94	14.76	21.87
NIGHT	GATED2GATED w/o passive	512x1024	Packnet(D=4)	✓	✓	✓	✗	✓	✓	9.97	0.25	5.03	**82.56**	**91.54**	94.25
NIGHT	GATED2GATED w/o masks	512x1024	Packnet(D=4)	✓	✓	✓	✓	✗	✗	10.05	0.27	5.36	80.06	90.44	93.75
NIGHT	GATED2GATED with infinity mask	512x1024	Packnet(D=4)	✓	✓	✓	✓	✗	✓	9.88	0.26	5.45	78.87	90.71	94.01
NIGHT	GATED2GATED with cycle mask	512x1024	Packnet(D=4)	✓	✓	✓	✓	✓	✗	9.58	0.25	5.03	80.68	91.25	94.40
NIGHT	**GATED2GATED** (final)	512x1024	Packnet(D=4)	✓	✓	✓	✓	✓	✓	**9.43**	**0.21**	**4.86**	82.17	**91.54**	**94.48**

Table 6. Ablation studies evaluated on the Gated2Depth test dataset. We investigate different depth net architectures, loss combinations, resolutions, and the impact of the presented masks on the network performance. Our final model outperforms all other methods by a significant margin.

		Method	clear					light fog					dense fog					snow				
			RMSE	MAE	δ_1	δ_2	δ_3	RMSE	MAE	δ_1	δ_2	δ_3	RMSE	MAE	δ_1	δ_2	δ_3	RMSE	MAE	δ_1	δ_2	δ_3
DAY	not binned	MONODEPTH RGB [5]	8.12	4.16	86.26	**94.94**	**97.40**	7.26	3.32	89.00	95.29	97.69	7.15	3.41	85.68	94.09	96.71	8.40	4.28	83.75	93.80	97.10
		SPARSE-TO-DENSE [12]	9.14	5.94	42.56	85.98	95.52	7.44	5.41	31.61	80.20	95.55	6.67	4.98	31.44	80.49	94.83	8.69	5.81	40.10	84.65	95.89
		PACKNET-SLIM RGB [10]	8.10	4.07	84.55	94.49	97.34	7.53	3.56	87.16	94.56	97.17	5.96	2.86	86.27	94.48	97.33	7.85	4.03	83.82	94.65	**97.62**
		PACKNET-SLIM G [10]	11.47	6.49	68.83	86.46	93.16	9.45	4.88	77.94	90.30	94.94	9.43	4.62	78.24	90.83	94.48	10.22	5.78	72.14	88.67	94.27
		MONODEPTH2 RGB [6]	12.45	8.13	44.62	76.21	91.14	9.71	6.42	50.28	79.94	89.98	7.79	5.45	48.11	76.12	85.86	11.46	7.56	45.61	77.81	92.06
		MONODEPTH2 G [6]	9.10	4.28	85.67	92.56	95.15	10.07	4.65	86.23	91.50	94.32	10.88	4.75	85.58	90.93	93.21	8.05	3.81	**86.47**	93.03	95.72
		GATED2DEPTH [8]	**6.67**	**3.30**	**87.36**	94.48	96.94	**4.69**	**2.39**	90.75	96.73	98.41	**4.39**	**2.51**	87.35	**96.54**	**98.05**	**6.61**	**3.39**	85.05	94.10	97.03
		GATED2GATED	7.12	3.62	87.22	94.89	97.02	4.89	2.49	**91.69**	**97.00**	**98.48**	6.25	2.87	**89.66**	95.58	97.02	7.01	3.75	84.74	**94.76**	97.11
	binned	MONODEPTH RGB [5]	12.74	8.43	75.11	90.18	94.81	14.04	9.10	72.70	88.43	94.32	14.67	10.64	63.49	82.90	91.89	13.17	8.73	71.56	89.06	94.81
		SPARSE-TO-DENSE [12]	13.66	9.85	54.20	82.42	91.47	14.23	10.66	49.75	79.62	90.10	18.50	15.35	37.04	64.67	78.25	13.42	9.81	53.12	82.29	92.04
		PACKNET-SLIM RGB [10]	12.76	8.93	70.34	89.34	**95.04**	14.36	9.80	68.30	87.85	94.51	11.74	7.66	76.63	90.07	**95.06**	12.57	8.75	69.89	89.93	**95.81**
		PACKNET-SLIM G [10]	16.46	11.62	56.91	78.43	88.48	16.95	11.80	59.09	78.80	88.81	17.01	12.09	54.93	76.37	88.89	15.30	10.33	62.22	82.29	90.65
		MONODEPTH2 RGB [6]	19.53	17.44	27.12	54.99	78.29	20.85	18.70	26.49	51.68	72.73	22.71	21.35	19.71	39.02	58.87	18.85	16.82	27.87	56.84	80.01
		MONODEPTH2 G [6]	13.26	7.40	78.59	88.95	92.77	18.17	10.43	72.91	83.20	89.27	15.56	8.72	76.79	85.38	90.68	12.84	7.12	**80.04**	89.34	93.13
		GATED2DEPTH [8]	11.48	6.60	79.17	87.38	91.58	11.28	6.63	81.20	88.66	92.56	11.86	7.85	71.72	87.10	91.70	11.28	6.61	78.87	87.93	92.50
		GATED2GATED	**11.15**	**6.31**	**80.82**	**90.48**	93.97	**10.70**	**6.01**	**84.71**	**91.52**	**94.65**	**11.09**	**6.86**	**81.09**	**91.43**	94.47	**10.97**	**6.28**	80.01	**91.12**	94.63
NIGHT	not binned	MONODEPTH RGB [5]	9.28	4.74	83.65	93.08	96.08	9.22	4.84	81.82	93.11	**96.62**	7.35	3.34	**90.60**	95.45	97.54	10.69	5.67	81.03	91.64	94.94
		SPARSE-TO-DENSE [12]	10.04	6.79	35.08	77.89	92.97	9.76	6.55	37.58	80.52	93.90	7.61	5.51	35.46	75.41	92.44	9.57	6.33	38.27	81.18	93.92
		PACKNET-SLIM RGB [10]	9.03	4.72	80.98	92.32	96.06	9.28	5.08	78.29	91.73	96.41	7.43	3.48	86.04	92.55	95.88	9.47	4.98	79.21	90.85	94.82
		PACKNET-SLIM G [10]	10.69	5.83	74.40	88.84	94.07	11.05	6.26	70.75	88.08	93.79	8.68	4.61	80.68	91.18	94.77	10.83	6.03	73.36	87.49	92.89
		MONODEPTH2 RGB [6]	13.94	9.03	43.84	71.48	84.97	13.58	8.82	44.20	73.14	87.48	10.06	6.90	47.09	70.65	80.59	13.75	8.96	42.08	71.86	86.63
		MONODEPTH2 G [6]	10.05	5.07	81.03	89.75	93.18	10.09	5.31	80.18	89.10	93.25	10.58	4.90	83.15	88.92	93.11	9.90	5.04	82.10	89.64	92.79
		GATED2DEPTH [8]	**6.08**	**2.89**	**89.89**	**95.02**	**96.92**	**6.73**	**3.60**	**83.83**	93.61	96.61	**4.58**	**2.58**	88.55	95.87	**97.90**	**6.89**	**3.72**	84.30	92.78	**95.70**
		GATED2GATED	7.30	3.83	86.40	94.48	96.58	7.27	4.00	83.67	**94.06**	96.44	5.63	3.09	88.63	**96.14**	**97.90**	8.20	4.26	**85.02**	**93.11**	95.51
	binned	MONODEPTH RGB [5]	13.78	8.92	72.63	88.48	93.37	13.30	8.75	72.52	88.88	**94.45**	16.31	10.99	69.15	85.92	**91.33**	15.28	9.89	68.88	86.86	**93.44**
		SPARSE-TO-DENSE [12]	14.43	10.40	50.32	78.66	89.58	13.92	10.01	51.88	80.41	90.70	16.54	12.07	47.15	73.45	84.36	14.08	10.05	52.91	81.03	90.85
		PACKNET-SLIM RGB [10]	13.71	9.60	66.75	86.20	92.99	13.99	10.34	61.44	84.63	93.92	14.78	10.75	63.86	82.00	90.28	14.37	9.91	65.36	86.33	93.37
		PACKNET-SLIM G [10]	15.81	11.11	59.80	79.52	88.65	16.01	11.23	58.44	80.60	89.57	17.49	12.60	57.72	77.77	87.26	16.47	11.40	60.17	79.55	88.62
		MONODEPTH2 RGB [6]	21.22	18.29	29.86	54.97	73.65	20.51	17.98	28.89	54.01	74.13	23.29	20.77	23.47	46.25	65.35	21.91	19.41	25.03	49.83	70.66
		MONODEPTH2 G [6]	14.52	8.30	74.45	85.41	89.73	14.21	8.29	74.56	85.06	91.01	18.33	11.88	66.61	79.25	84.48	15.11	8.46	76.15	86.38	90.52
		GATED2DEPTH [8]	**10.06**	**5.17**	**84.81**	**90.59**	**93.39**	**9.94**	**5.37**	**81.95**	**89.80**	93.63	**12.51**	**7.72**	**76.90**	86.59	90.81	**10.70**	**5.81**	**81.81**	89.45	93.02
		GATED2GATED	11.69	6.74	80.25	89.58	92.83	11.29	6.46	79.39	89.31	93.17	13.52	8.69	76.43	**86.70**	90.61	11.91	6.80	80.76	**90.09**	93.31

Table 7. Quantitative results of the proposed Gated2Gated framework and state-of-the-art-methods for the Seeing Through Fog [1] dataset. All metrics are also evaluated for bins of approximately 7m to weight all distances equally.

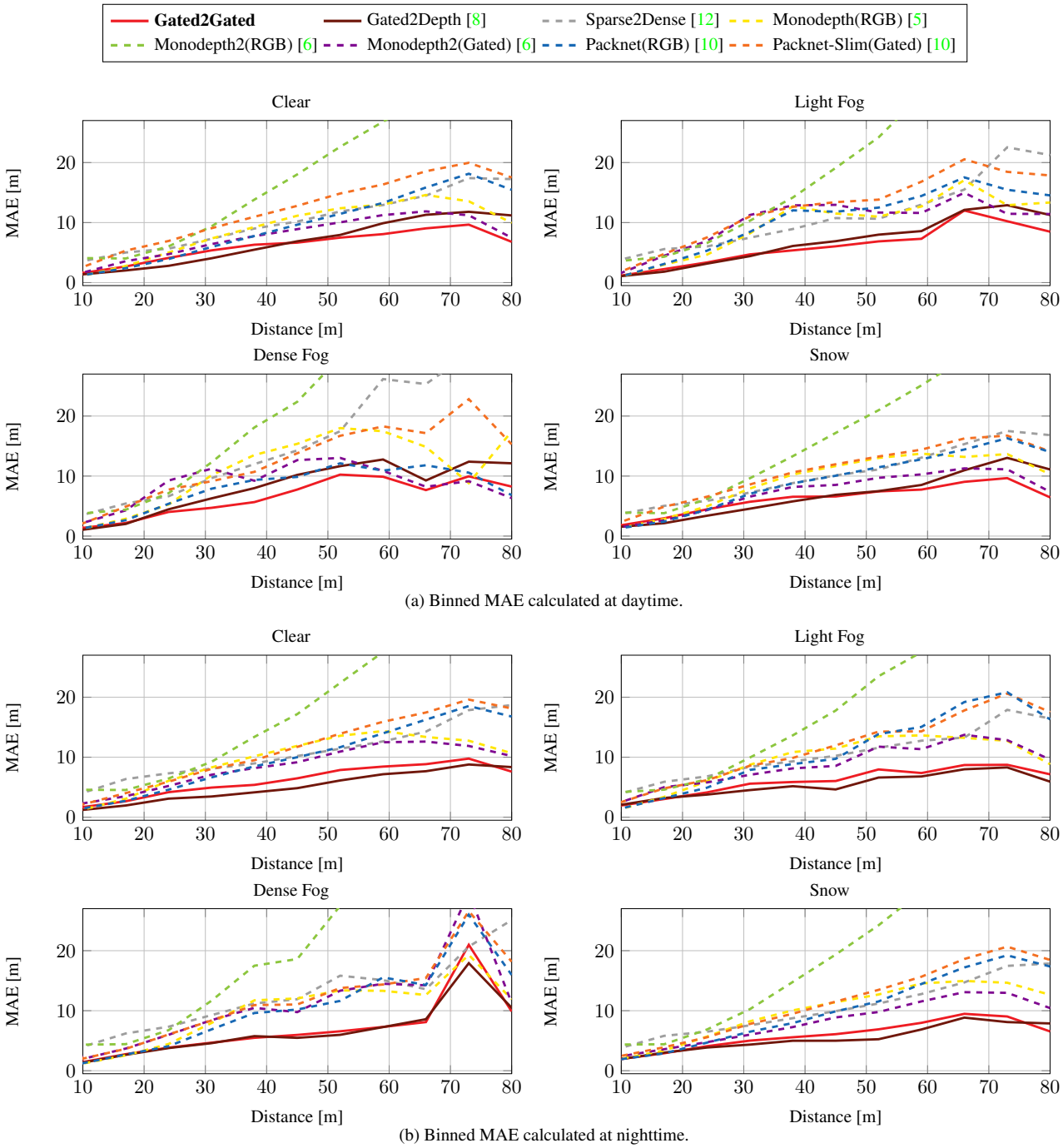
night time conditions, our method performs on par with Gated2Depth [8]. For completeness, we show the non-binned depth evaluation scores in Table 7.

Note that, to obtain a clear ground truth in adverse weather, we used the DROR algorithm [2] according to the author's specifications. This helped remove cluttered points from the LiDAR point clouds providing noise-free ground truth for evaluation in challenging adverse weather scenarios. With this approach, we removed about 8.2 % of the LiDAR points.

9. Additional Qualitative Result

Next, we present additional qualitative results of the proposed Gated2Gated method and other state-of-the-art methods. Depth map predictions of the different approaches, corresponding RGB and gated images as well as the LiDAR measurements are shown in Figures 8, 9, 10, 11, 12, 13, 14, 15, and 16. These results further validate the improved performance of our method in capturing structural details of the objects even at far distances. For example, in Figure 8 our method is able to show distinct object contours for the pedestrians holding hands. Our method is able to distinguish depth of very fine structures in the scene from background e.g, legs of the person on right side (Figure 8). We find that only Packnet model trained on gated images is able to achieve results coming close to our method. However, this method often fails estimating the correct depth values for moving objects (see Figure 12 and 16). The daytime results in Figure 9, 10, 12, 13, 15, and 16. indicate that our method also works well in the presence of sunlight and is able to handle strong ambient illumination. We have also compared our proposed method for different weather conditions with Gated2Depth [8] and LiDAR point clouds. Our proposed method also works well under low light conditions as apparent from Figure 19.

For fog and snow conditions (Figure 20 & 21), our proposed method generates robust depth predictions as compared to Gated2Depth which fails for thick fog and heavy snow. Interestingly, even unfiltered LiDAR point clouds fail for these adverse weather conditions, whereas our proposed method still generates good qualitative results. This highlights the robustness of our approach.



(a) Binned MAE calculated at daytime.

(b) Binned MAE calculated at nighttime.

Figure 7. MAE calculated over depth bins of approximately 7 m in different weather conditions. **Gated2Gated** outperforms all other methods especially at nighttime and for far distances.

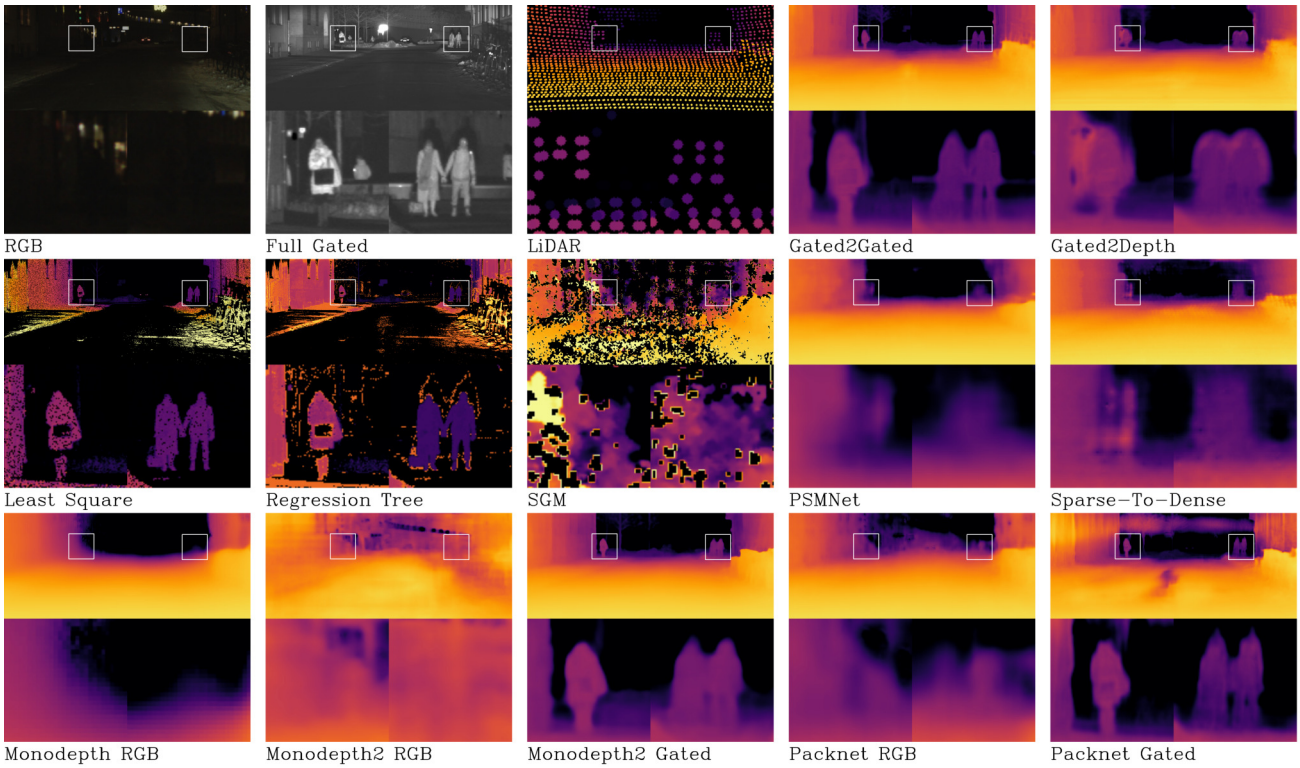


Figure 8. Qualitative comparison of the proposed Gated2Gated approach and state-of-the-art methods. For each example, we show the corresponding RGB image, the full gated image and the LiDAR measurements. Gated2Gated predicts finer grain details and sharper object contours in the depth maps than the other approaches.

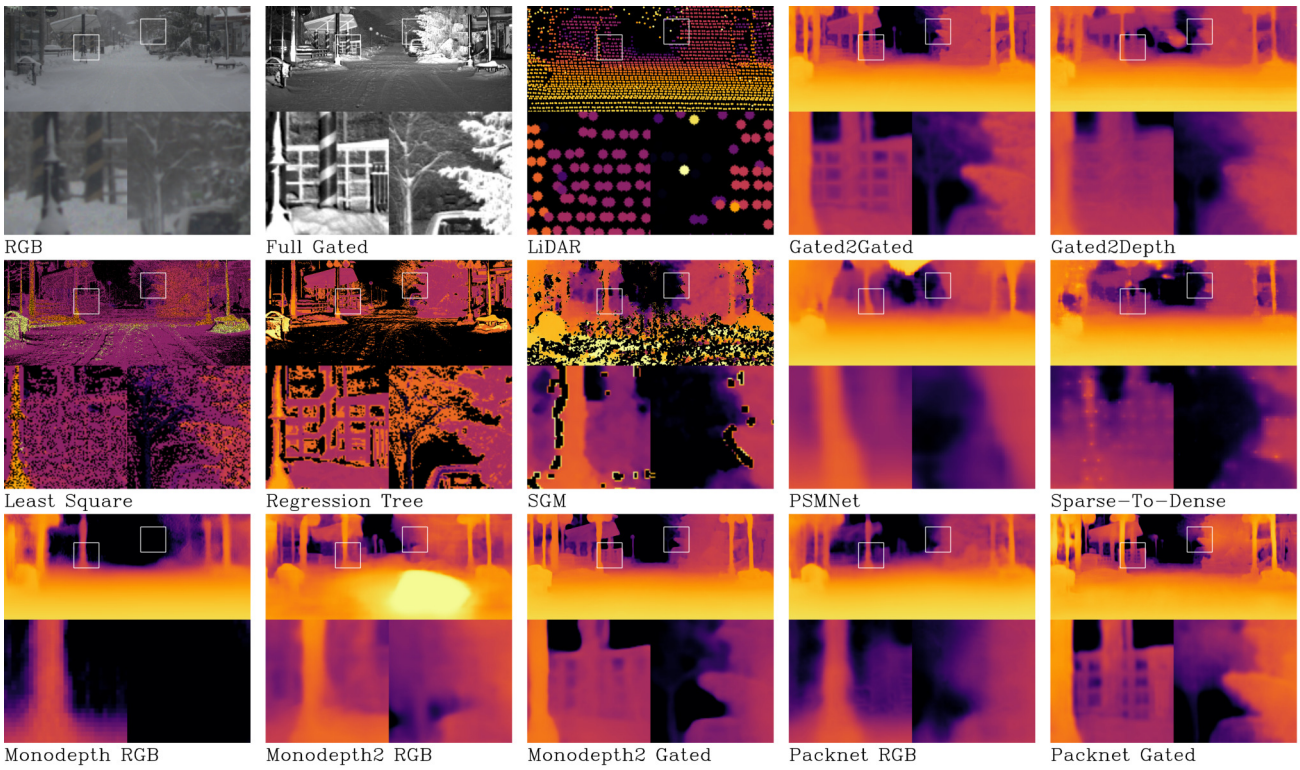


Figure 9. Additional qualitative comparison of our Gated2Gated approach and state-of-the-art methods.

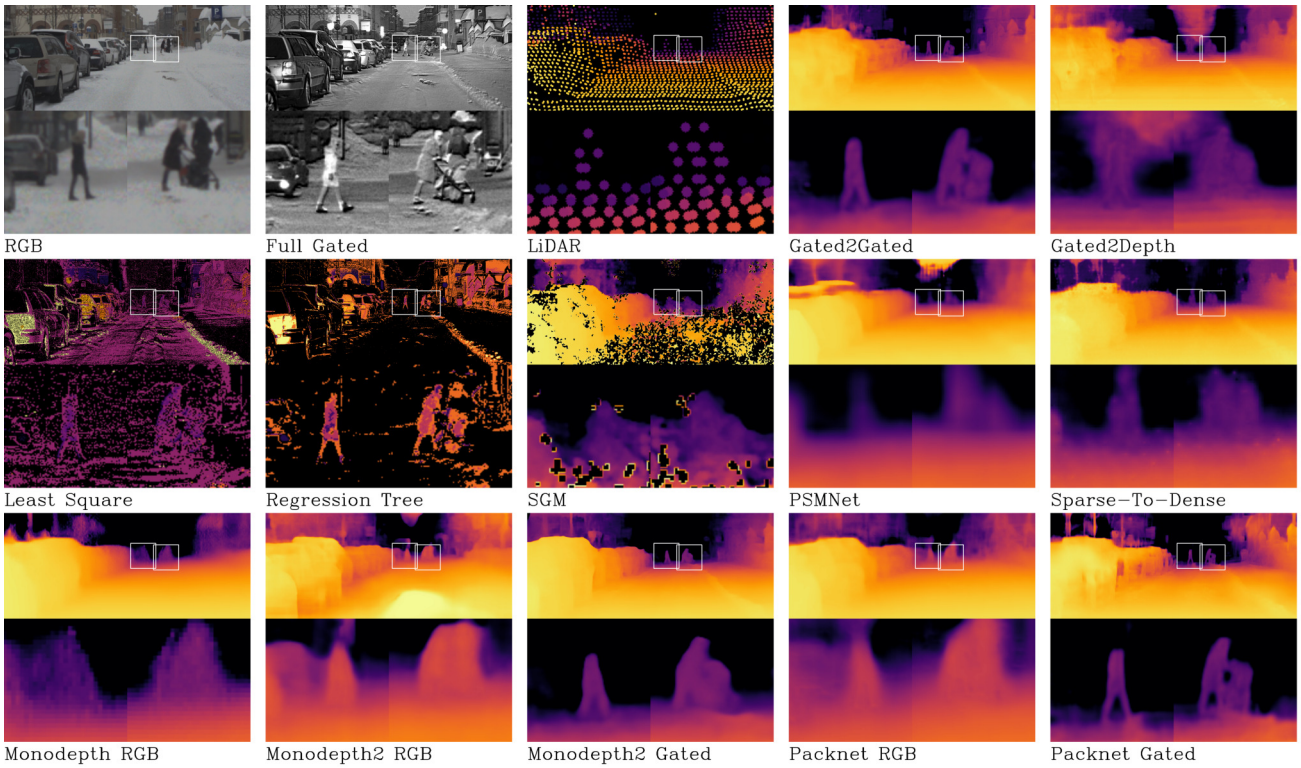


Figure 10. Additional qualitative comparison of our Gated2Gated approach and state-of-the-art methods.

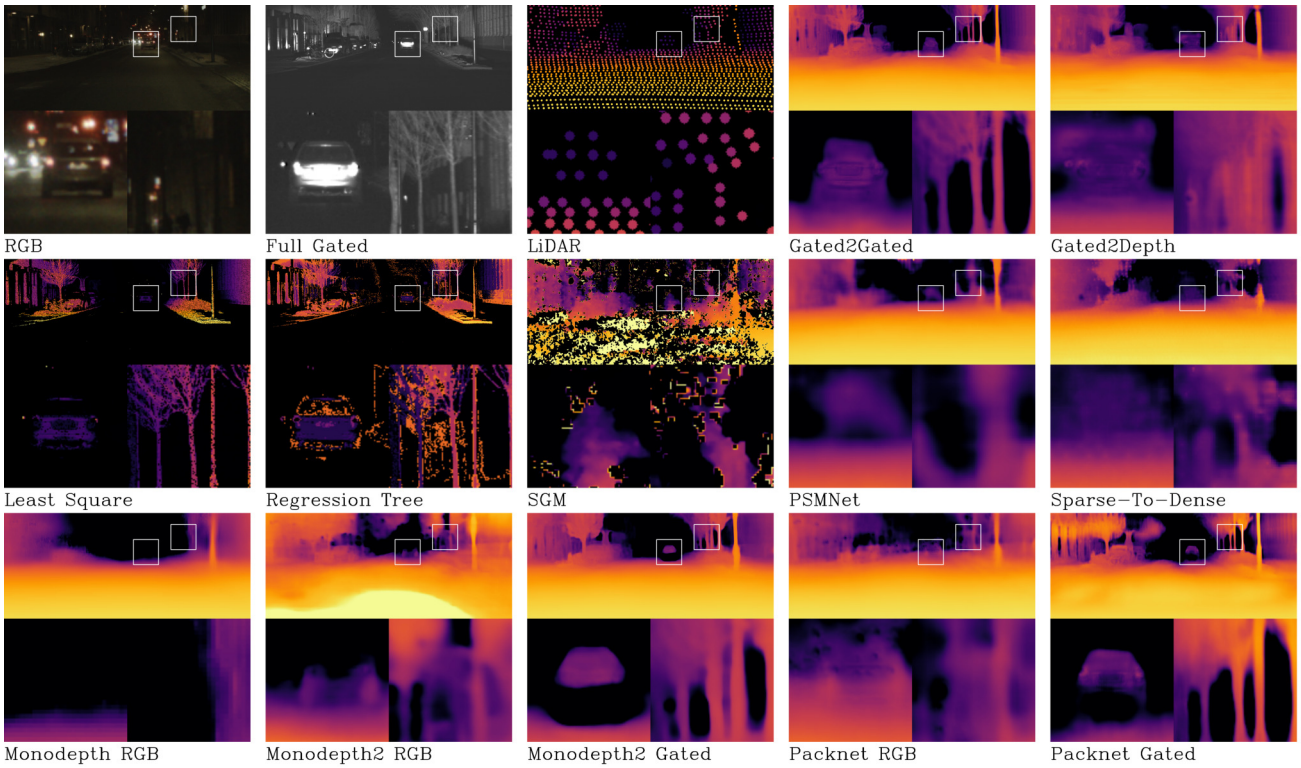


Figure 11. Additional qualitative comparison of our Gated2Gated approach and state-of-the-art methods.

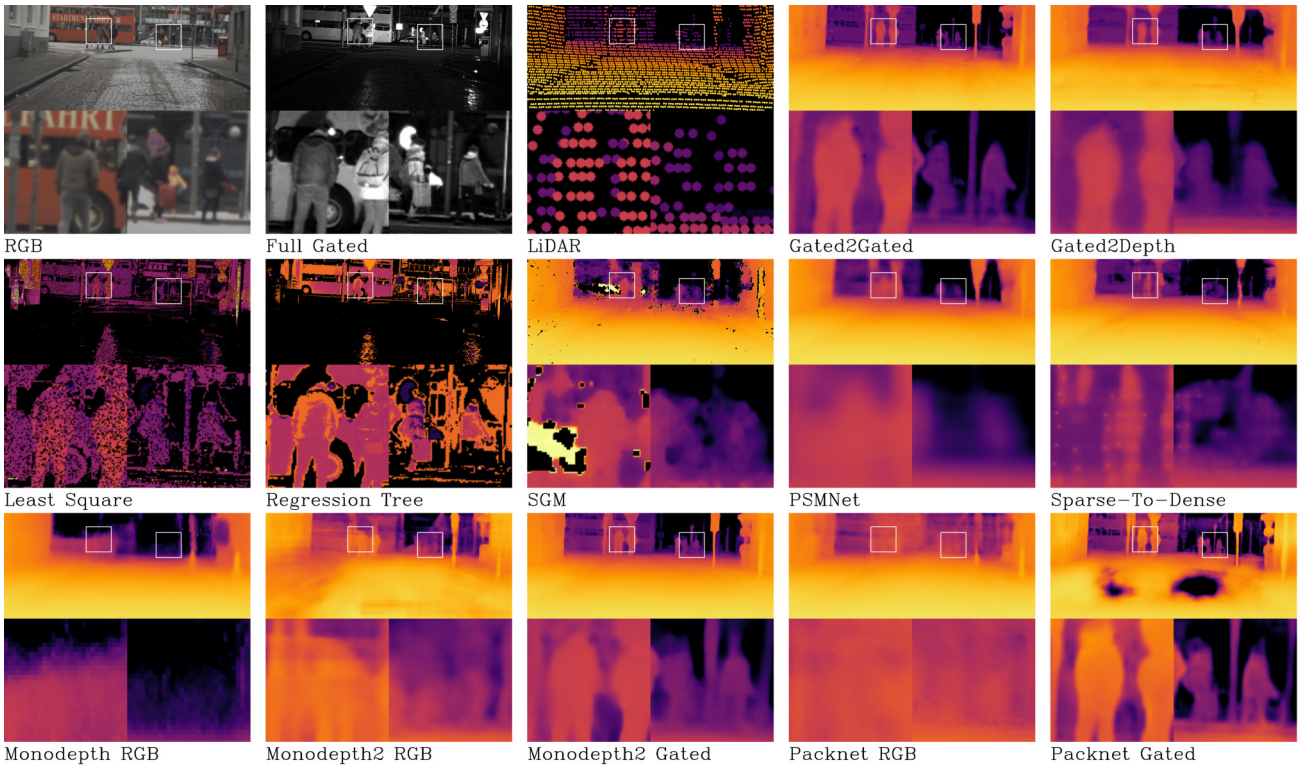


Figure 12. Additional qualitative comparison of our Gated2Gated approach and state-of-the-art methods.

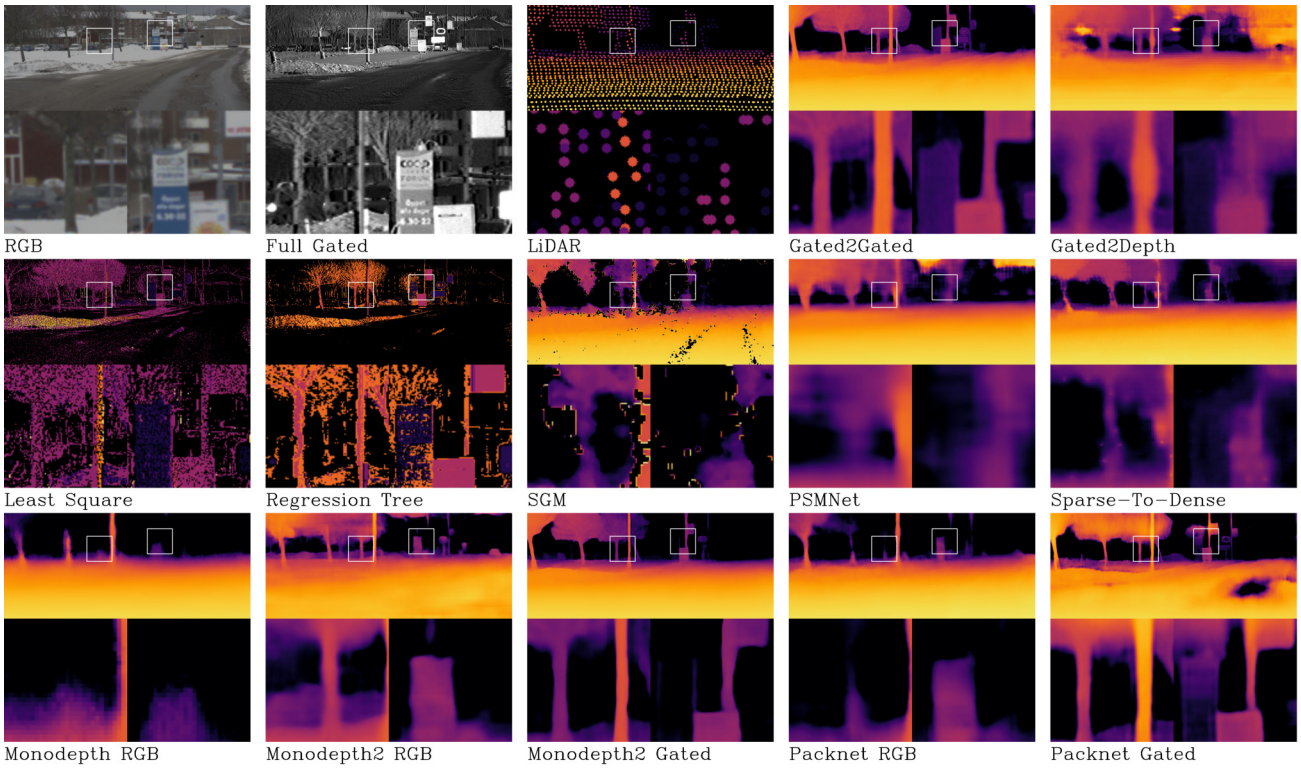


Figure 13. Additional qualitative comparison of our Gated2Gated approach and state-of-the-art methods.

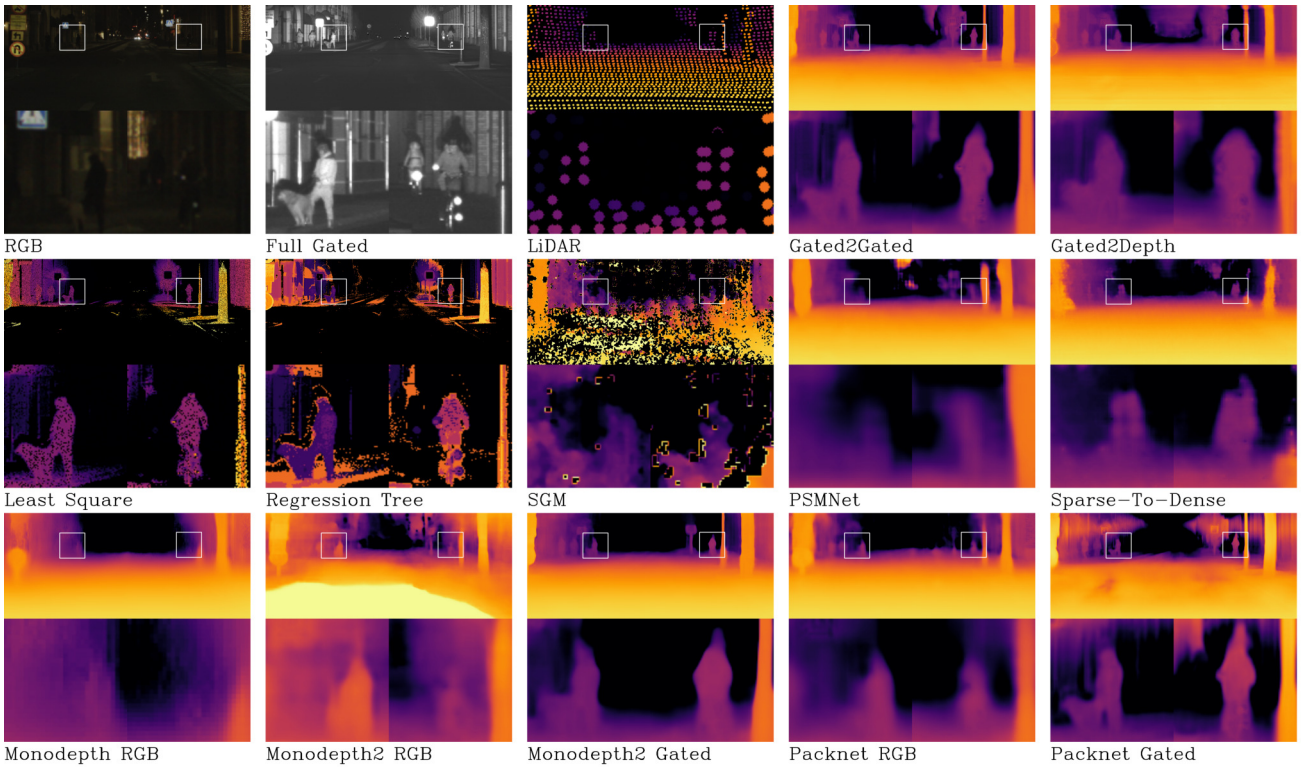


Figure 14. Additional qualitative comparison of our Gated2Gated approach and state-of-the-art methods.

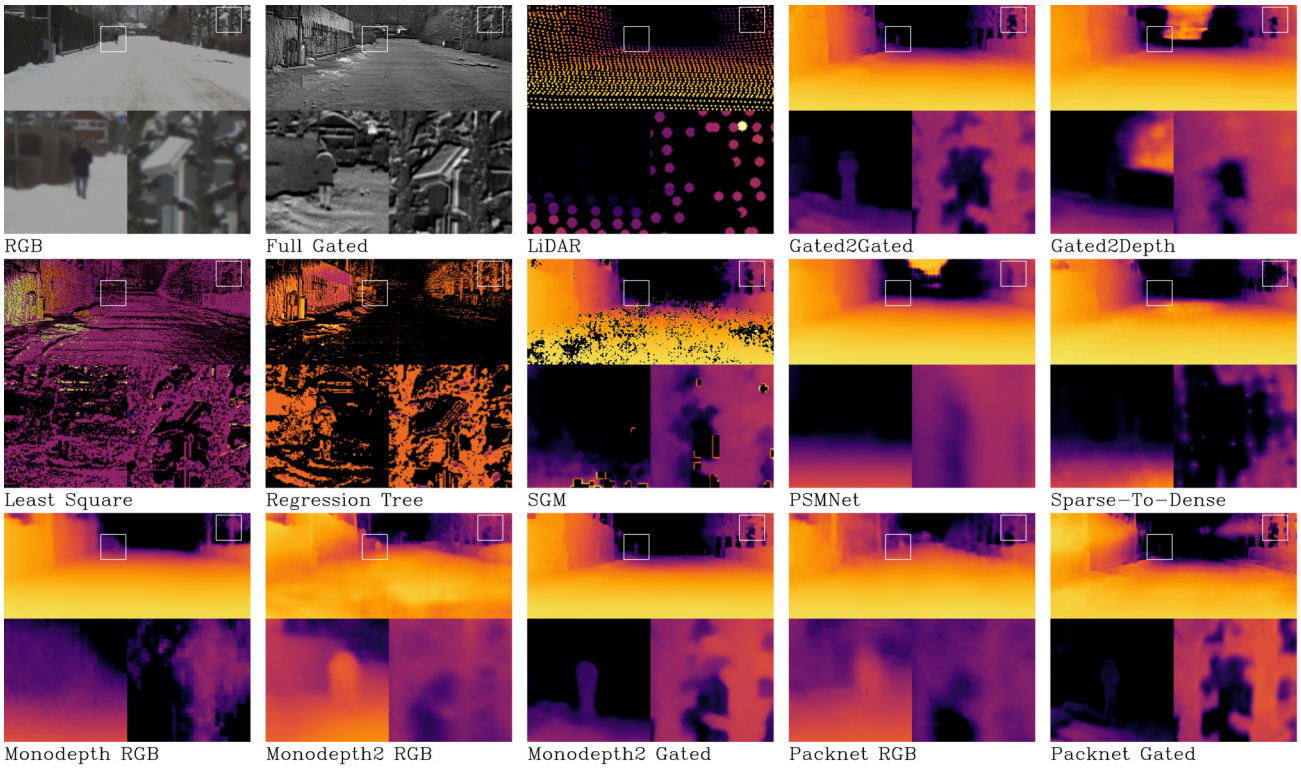


Figure 15. Additional qualitative comparison of our Gated2Gated approach and state-of-the-art methods.

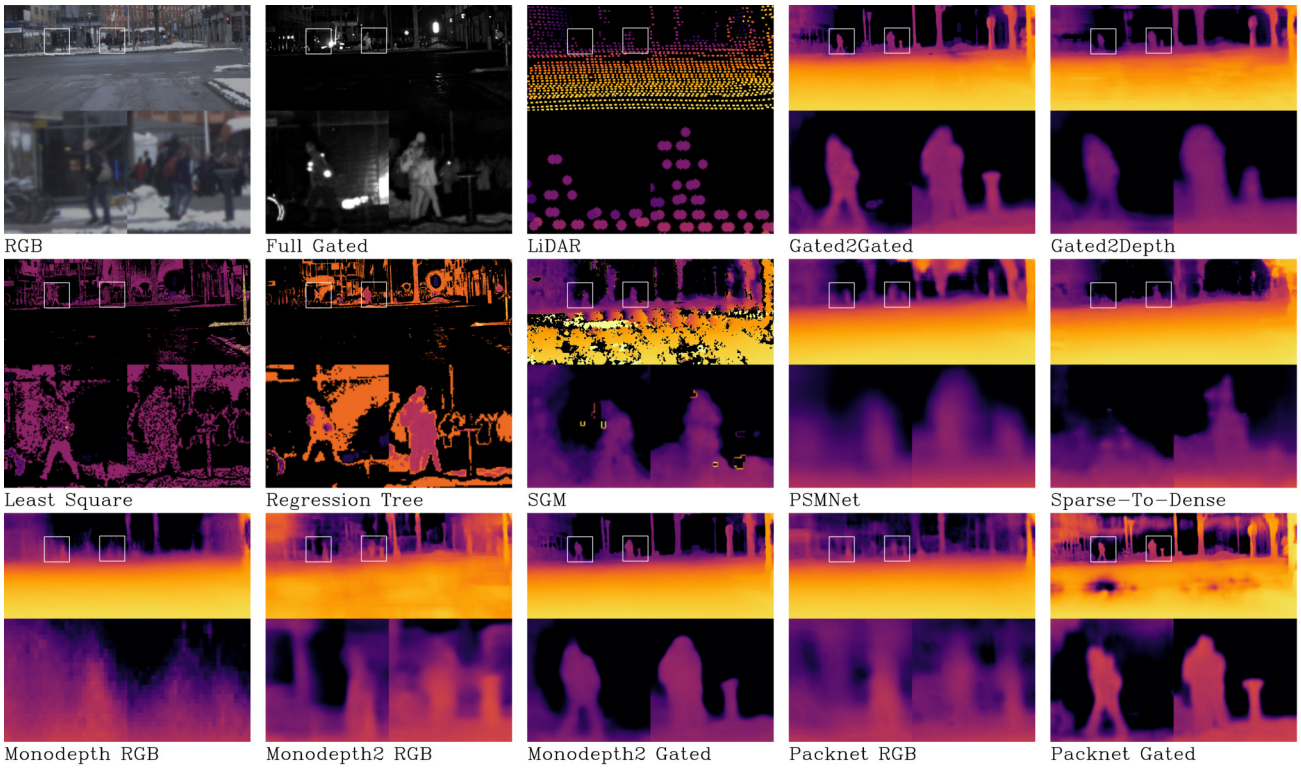


Figure 16. Additional qualitative comparison of our Gated2Gated approach and state-of-the-art methods.

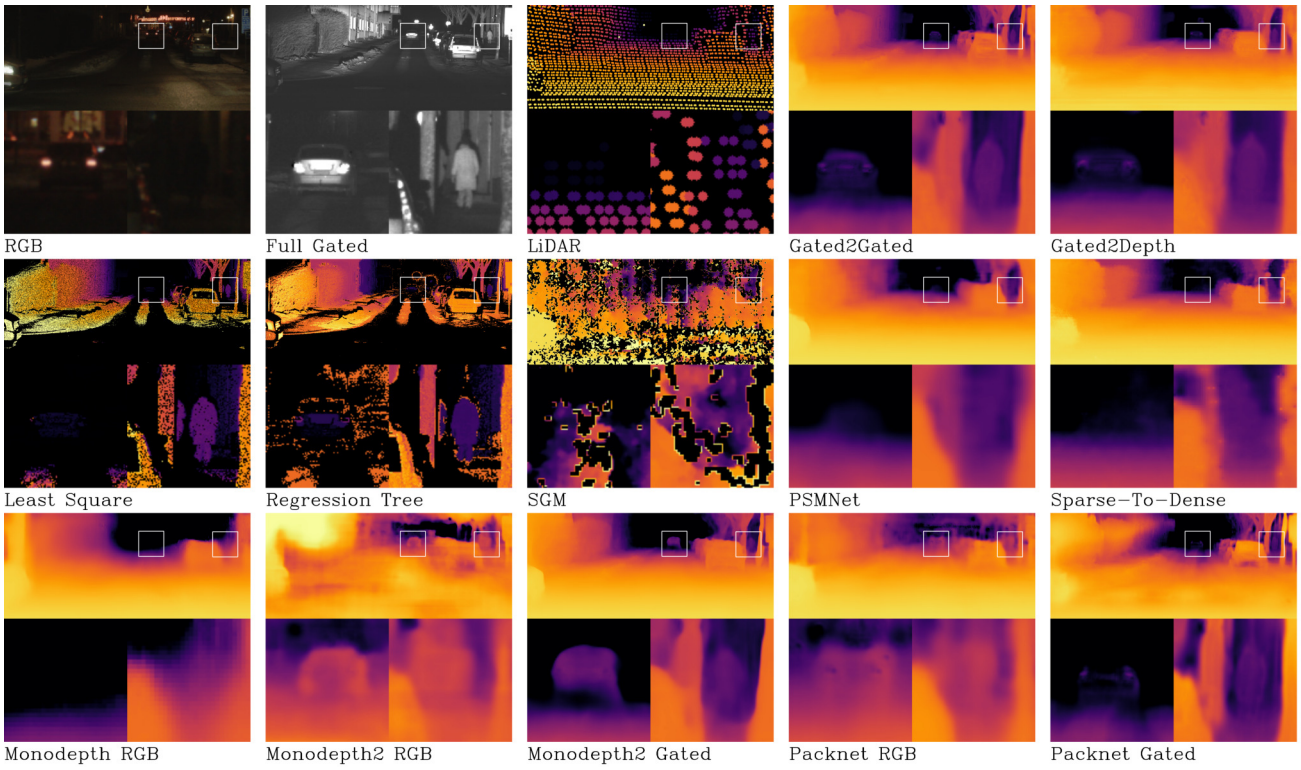


Figure 17. Additional qualitative comparison of our Gated2Gated approach and state-of-the-art methods.

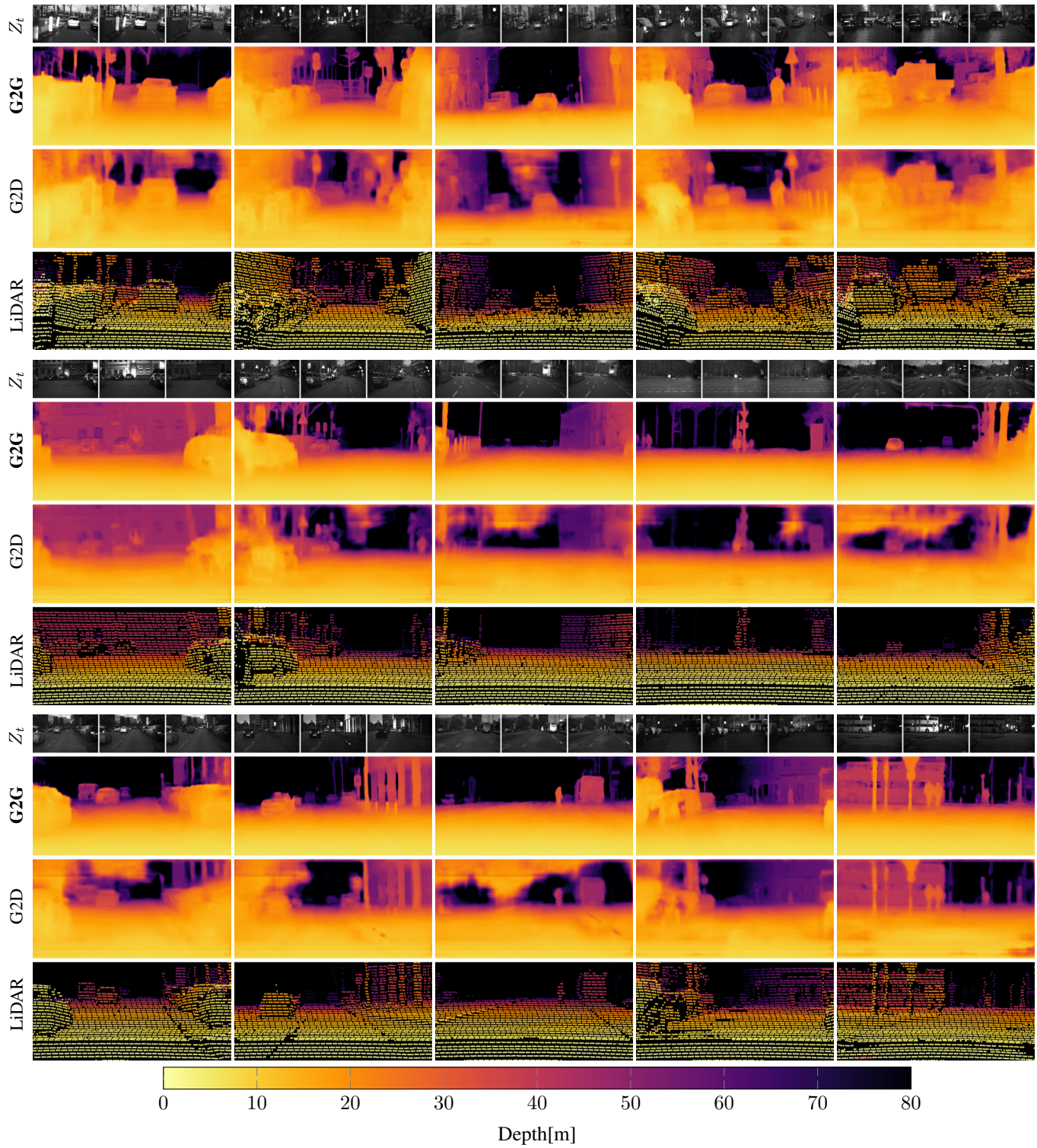


Figure 18. Qualitative comparison of **Gated2Gated**(G2G) with Gated2Depth [8] (supervised) for **Clear Day** conditions. It is evident from these examples that our proposed method is robust to strong ambient light whereas Gated2Depth fails in such scenarios.

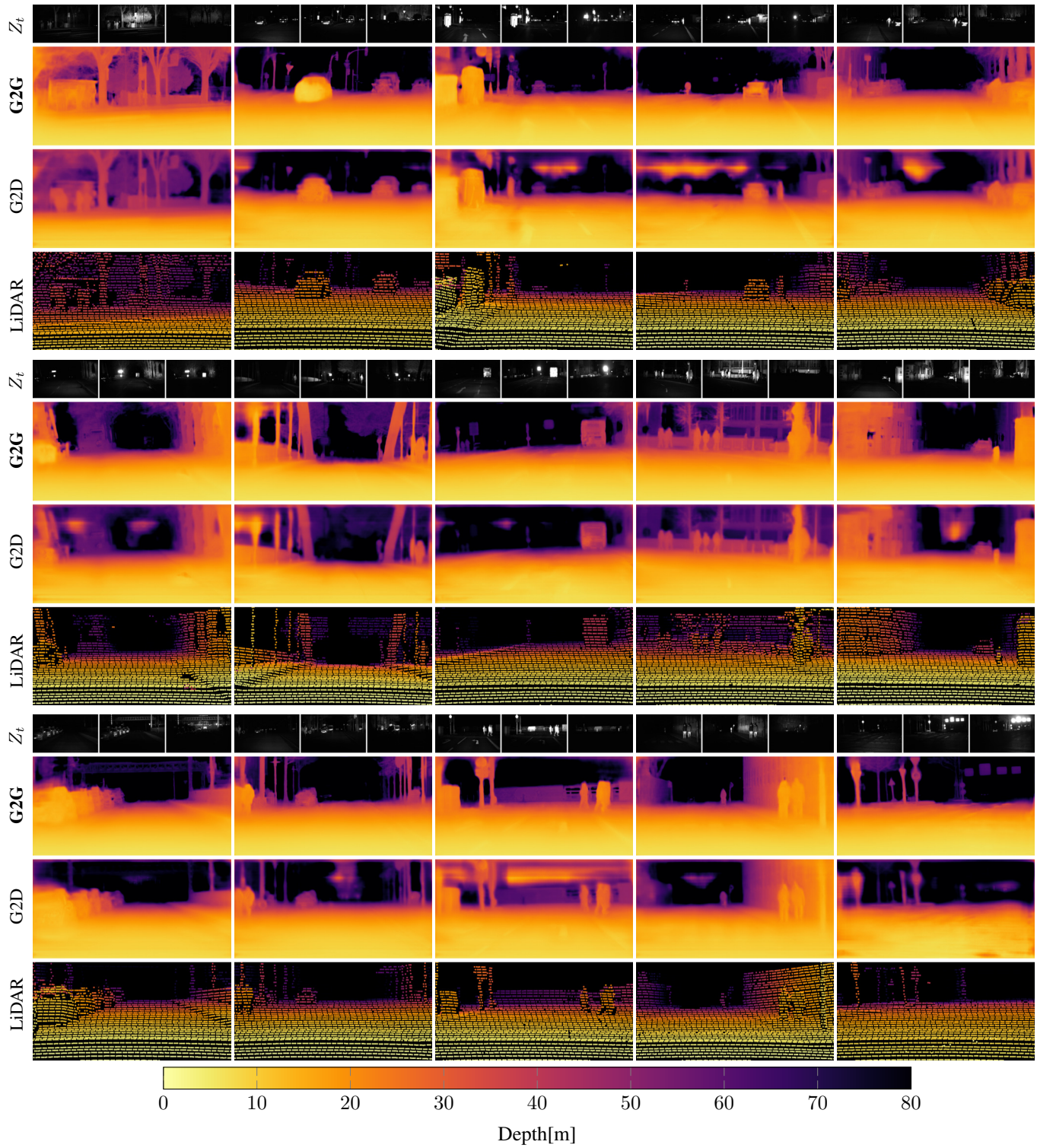


Figure 19. Qualitative comparison for **Clear Night** conditions. Our proposed method performs better than Gated2Depth even in low-light conditions, especially mitigating the artefacts for large depths in Gated2Depth.

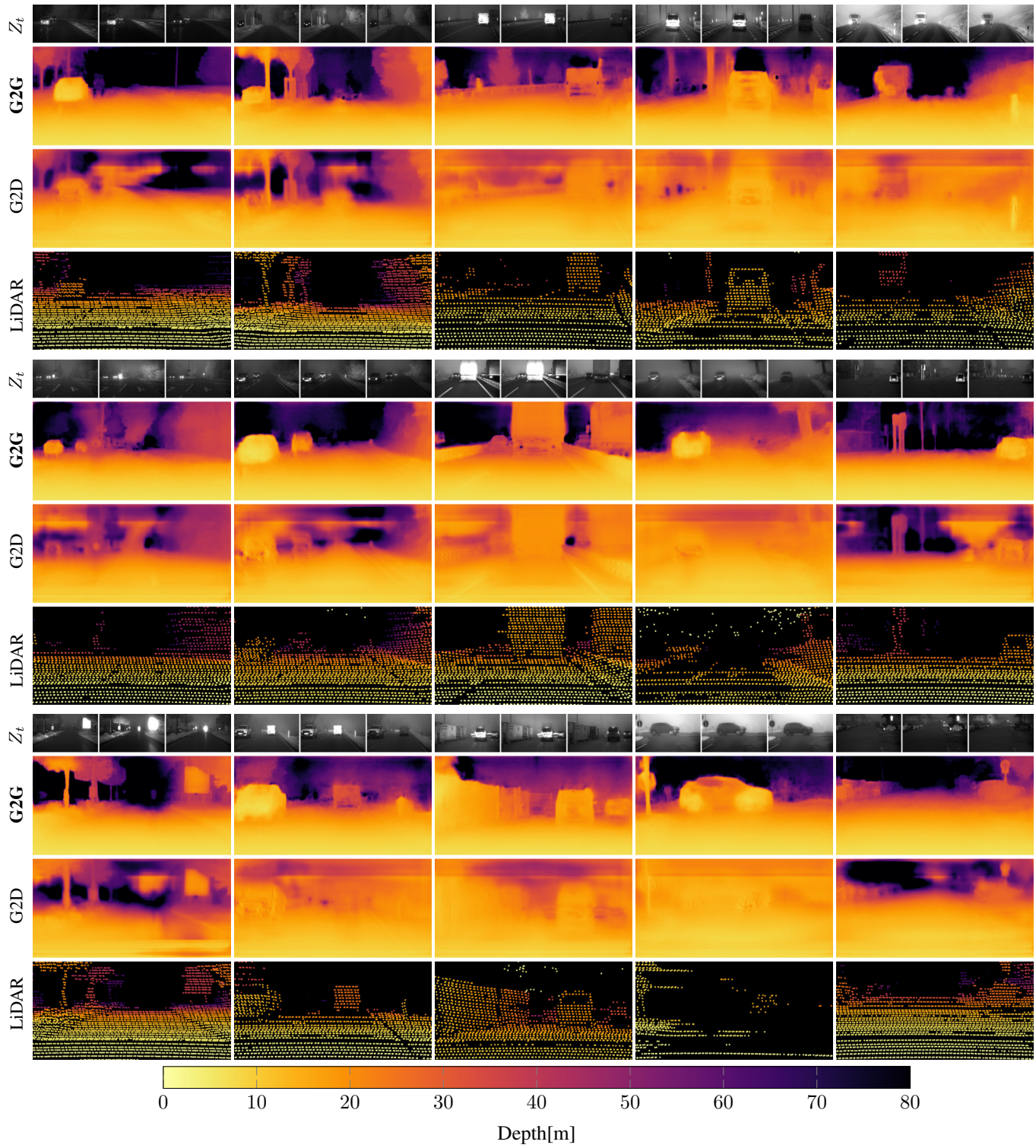


Figure 20. Qualitative comparison for **Fog** conditions. Our proposed method performs robustly in scattering medium for which both LiDAR and Gated2Depth suffer to the point of failure.

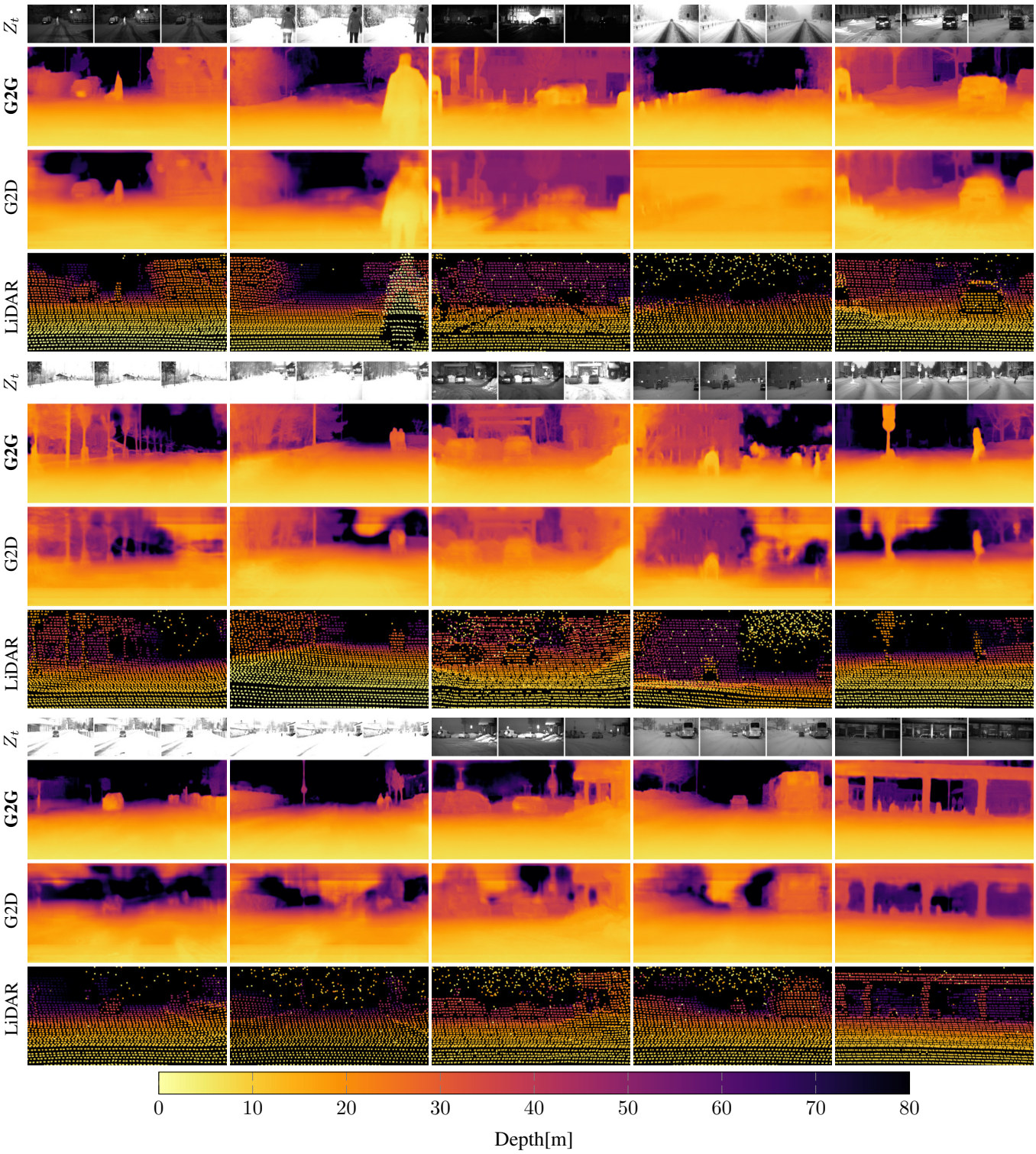


Figure 21. Qualitative comparison for **Snow** weather conditions. Note that the proposed method performs better than Gated2Depth [9] mitigating the artefacts in depth due to scattering by snow.

References

[1] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 8, 10

[2] Nicholas Charron, Stephen Phillips, and Steven L. Waslander. De-noising of lidar point clouds corrupted by snowfall. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 254–261, 2018. 10

[3] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. 2

[4] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 7

[5] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 10, 11

[6] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 7, 8, 10, 11

[7] Tobias Gruber, Mario Bijelic, Felix Heide, Werner Ritter, and Klaus Dietmayer. Pixel-accurate depth evaluation in realistic driving scenarios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 95–105. IEEE, 2019. 8

[8] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time dense lidar from gated images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 10, 11, 17

[9] Tobias Gruber, Mariia Kokhova, Werner Ritter, Norbert Haala, and Klaus Dictmayer. Learning super-resolved depth from active gated imaging. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3051–3058. IEEE, 2018. 7, 20

[10] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020. 4, 7, 8, 10, 11

[11] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *IEEE International Conference on Robotics and Automation*, 2019. 7

[12] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *IEEE International Conference on Robotics and Automation*, pages 1–8, 2018. 10, 11

[13] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4